Maximum likelihood estimation in the logistic regression model with a cure fraction

Aba Diop, Aliou Diop, Jean-François Dupuy

Maximum likelihood estimation in the logistic regression model with a cure fraction

Aba Diop

Laboratoire Mathématiques, Image et Applications, Université de La Rochelle, Avenue Michel Crépeau, La Rochelle, France, and Laboratoire d'Etudes et de Recherches en Statistiques et Développement, Université Gaston Berger, Saint Louis, Sénégal
Email: aba.diop@univ-lr.fr

Aliou Diop

Laboratoire d'Etudes et de Recherches en Statistiques et Développement, Université Gaston Berger, Saint Louis, Sénégal.
Email: alioudiop52@yahoo.fr

Jean-François Dupuy†

Laboratoire Mathématiques, Image et Applications, Université de La Rochelle, Avenue Michel Crépeau, La Rochelle, France
Email: jean-francois.dupuy@univ-lr.fr

Abstract. Logistic regression is widely used in medical studies to investigate the relationship between a binary response variable Y and a set of potential predictors $\mathbf{X}$. The binary response may represent, for example, the occurrence of some outcome of interest ($Y = 1$ if the outcome occurred and $Y = 0$ otherwise). In this paper, we consider the problem of estimating the logistic regression model with a cure fraction. A sample of observations is said to contain a cure fraction when a proportion of the study subjects (the so-called cured individuals, as opposed to the susceptibles) cannot experience the outcome of interest. One problem arising then is that it is usually unknown who are the cured and the susceptible subjects, unless the outcome of interest has been observed. In this setting, a logistic regression analysis of the relationship between $\mathbf{X}$ and Y among the susceptibles is no more straightforward. We develop a maximum likelihood estimation procedure for this problem, based on the joint modeling of the binary response of interest and the cure status. We investigate the identifiability of the resulting model. Then, we establish the consistency and asymptotic normality of the proposed estimator, and we conduct a simulation study to investigate its finite-sample behavior.

Keywords: Zero-inflation, Maximum likelihood estimation, Consistency, Asymptotic normality, Simulations

1. Introduction

Logistic regression is widely used to model binary response data in medical studies. An example of a binary response variable is the infection status (infected *vs* uninfected) with respect to some disease. A logistic regression model can be used to investigate the relationship between the infection status and various potential predictors. If Y_i denotes the infection status for the i-th individual in a sample of size n ($Y_i = 1$ if the individual is

†Corresponding author

infected, and $Y_i = 0$ otherwise), and $\mathbf{X}_i$ denotes the corresponding (p-dimensional, say) predictor, the logistic regression model expresses the relationship between Y_i and $\mathbf{X}_i$ in term of the conditional probability $\mathbb{P}(Y_i = 1|\mathbf{X}_i)$ of infection, as:

$$\log\left(\frac{\mathbb{P}(Y_i = 1|\mathbf{X}_i)}{1 - \mathbb{P}(Y_i = 1|\mathbf{X}_i)}\right) = \beta'\mathbf{X}_i,$$

where $\beta \in \mathbb{R}^p$ is an unknown parameter to be estimated. An extensive literature has been devoted so far to statistical inference in logistic regression models. Estimation and testing procedures for this class of models are now well established and are available in standard statistical softwares. In particular, the maximum likelihood estimator of β is obtained by solving the following score equation:

$$\sum_{i=1}^{n} \mathbf{X}_i \left(Y_i - \frac{e^{\beta'\mathbf{X}_i}}{1 + e^{\beta'\mathbf{X}_i}}\right) = 0.$$

Asymptotic results (consistency and asymptotic normality) for this estimator were given by Gouriéroux and Monfort (1981) and Fahrmeir and Kaufmann (1985), among others. We refer the reader to Hosmer and Lemeshow (2000) and Hilbe (2009) for detailed treatments and numerous examples.

In this paper, we consider the problem of estimation in the logistic regression model with a cure fraction. In medical studies, it often arises that a proportion of the study subjects cannot experience the outcome of interest. Such individuals are said to be cured, or immune. The population under study can then be considered as a mixture of cured and susceptible subjects, where a subject is said to be susceptible if he would eventually experience the outcome of interest. One problem arising in this setting is that it is usually unknown who are the susceptible, and the cured subjects (unless the outcome of interest has been observed). Consider, for example, the occurrence of infection from some disease to be the outcome of interest. Then, if a subject is uninfected, the investigator usually does not know whether this subject is immune to the infection, or susceptible albeit still uninfected.

Estimating a regression model with a cure fraction can be viewed as a zero-inflated regression problem. Zero-inflation occurs in the analysis of count data when the observations contain more zeros than expected. Failure to account for these extra zeros is known to result in biased parameter estimates and inferences. The regression analysis of count data with excess zeros has attracted much attention so far. For example, Lambert (1992) proposed the zero-inflated Poisson (ZIP) regression model for count data with many zeros. This was further extended to a semiparametric ZIP regression model by Lam et al. (2006). We refer to Dietz and Böhning (2000) and Xiang et al. (2007) for a review of various other extensions of the ZIP model. Other popular models are the zero-inflated binomial (ZIB) regression model (see, for example, Hall (2000)), and the zero-inflated negative binomial (ZINB) regression model (see, for example, Ridout et al. (2001)). Recently, Kelley and Anderson (2008) proposed a zero-inflated proportional odds model (ZIPO) for ordinal outcomes, when some individuals are not susceptible to the phenomenon being measured. Various other models and numerous references can be found in Famoye and Singh (2006) and Lee et al. (2006).

In our paper, we consider the problem of estimating a logistic regression model from binary response data with a cure fraction, when the cure probability is modeled by a logistic regression. This can be viewed as a zero-inflated Bernoulli regression problem, where

logistic link functions are used for both the binary response of interest (the probability of infection, say) and the zero-inflation probability (the probability of being cured). The literature on zero-inflated models is extensive but to the best of our knowledge, the theoretical and numerical issues related to the statistical inference in this model have not been yet investigated. In this paper, we intend to fill this gap. We first investigate the identifiability question in this model. Then, we turn to the problem of estimation. The estimator we propose is obtained by maximizing the joint likelihood for the binary response of interest and the cure indicator. We prove the almost sure asymptotic existence, the consistency, and the asymptotic normality of this estimator. Then, we investigate its finite-sample properties via simulations.

The rest of this paper is organized as follows. In Section 2, we describe the problem of logistic regression with a cure fraction, and we propose an estimation method adapted to this setting. The proposed procedure is based on a joint regression model for the binary response of interest and the cure indicator. In Section 3, we investigate the identifiability of this model, and we state some regularity conditions. In Section 4, we derive the asymptotic properties of the resulting estimator. Section 5 describes a simulation study, where we numerically investigate the small to large sample properties of this estimator. A real data example illustrates the methodology. A discussion and some perspectives are given in Section 6.

2. Logistic regression with a cure fraction

2.1. Notations and the model set-up

Let $(Y_1, S_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, S_n, \mathbf{X}_n, \mathbf{Z}_n)$ be independent and identically distributed copies of the random vector $(Y, S, \mathbf{X}, \mathbf{Z})$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For every individual $i = 1, \ldots, n$, Y_i is a binary response variable indicating say, the infection status with respect to some disease (that is, $Y_i = 1$ if the i-th individual is infected, and $Y_i = 0$ otherwise), and S_i is a binary variable indicating whether individual i is susceptible to the infection ($S_i = 1$) or immune ($S_i = 0$). If $Y_i = 0$, then the value of S_i is unknown. Let $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})'$ and $\mathbf{Z}_i = (1, Z_{i2}, \ldots, Z_{iq})'$ be random vectors of predictors or covariates (both categorical and continuous predictors are allowed). We shall assume in the following that the $\mathbf{X}_i$'s are related to the infection status, while the $\mathbf{Z}_i$'s are related to immunity. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are allowed to share some components.

The logistic regression model for the infection status assumes that the conditional probability $\mathbb{P}(Y = 1|\mathbf{X}_i, S_i)$ of infection is given by

$$\log\left(\frac{\mathbb{P}(Y = 1|\mathbf{X}_i, S_i)}{1 - \mathbb{P}(Y = 1|\mathbf{X}_i, S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} := \beta'\mathbf{X}_i \tag{1}$$

if $\{S_i = 1\}$, and by

$$\mathbb{P}(Y = 1|\mathbf{X}_i, S_i) = 0 \tag{2}$$

if $\{S_i = 0\}$, where $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown regression parameter measuring the association between potential predictors and the risk of infection (for a susceptible individual).

The statistical analysis of infection data with model (1) includes estimation and testing for β. Without immunity (that is, if $S_i = 1$ for every $i = 1, \ldots, n$), inference on β from the sample $(Y_1, \mathbf{X}_1, \mathbf{Z}_1), \ldots, (Y_n, \mathbf{X}_n, \mathbf{Z}_n)$ can be based on the maximum likelihood principle. When immunity is present, deriving the maximum likelihood estimator of β is no longer

straightforward: if $Y_i = 0$, we do not know whether $\{S_i = 1\}$, so that (1) applies, or whether $\{S_i = 0\}$, so that (2) applies.

One solution is to consider every individual i such that $\{Y_i = 0\}$ as being susceptible that is, to ignore a possible immunity of this individual. We may however expect this method to produce biased estimates of the association of interest (such a method will be evaluated in the simulation study described in section 5). Therefore in this paper, we aim at providing an alternative estimation procedure for β. This can be achieved if a model for immunity is available, as is explained in the next section.

2.2. The proposed estimation procedure

A model for the immunity status is defined through the conditional probability $\mathbb{P}(S = 1|\mathbf{Z}_i)$ of being susceptible to the infection. A common choice for this is the logistic model (see, for example, Fang et al. (2005) and Lu (2008, 2010) who considered estimation in various survival regression models with a cure fraction):

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2} + \ldots + \theta_q Z_{iq} := \theta'\mathbf{Z}_i \tag{3}$$

where $\theta = (\theta_1, \ldots, \theta_q)' \in \mathbb{R}^q$ is an unknown regression parameter.

Remark 1. We note that the model defined by (1)-(2)-(3) can be viewed as a zero-inflated Bernoulli regression model, with logit links for both the binary response of interest and the zero-inflation component. As far as we know, no theoretical investigation of this model has been undertaken yet. Such a work is carried out in the following.

From (1), (2), and (3), a straightforward calculation yields that

$$\mathbb{P}(Y=1|\mathbf{X}_i,\mathbf{Z}_i) = \frac{e^{\beta'\mathbf{X}_i+\theta'\mathbf{Z}_i}}{(1+e^{\beta'\mathbf{X}_i})(1+e^{\theta'\mathbf{Z}_i})}.$$

Let $\psi := (\beta', \theta')'$ denote the unknown k-dimensional ($k = p+q$) parameter in the conditional distribution of Y given $\mathbf{X}_i$ and $\mathbf{Z}_i$. ψ includes both β (considered as the parameter of interest) and θ (considered as a nuisance parameter). Now, the likelihood for ψ from the independent sample $(Y_i, S_i, \mathbf{X}_i, \mathbf{Z}_i)$ $(i = 1, \ldots, n)$ (where S_i is unknown when $Y_i = 0$) is as follows:

$$L_n(\psi) = \prod_{i=1}^{n}\left\{\left[\frac{e^{\beta'\mathbf{X}_i+\theta'\mathbf{Z}_i}}{(1+e^{\beta'\mathbf{X}_i})(1+e^{\theta'\mathbf{Z}_i})}\right]^{Y_i}\left[1-\frac{e^{\beta'\mathbf{X}_i+\theta'\mathbf{Z}_i}}{(1+e^{\beta'\mathbf{X}_i})(1+e^{\theta'\mathbf{Z}_i})}\right]^{1-Y_i}\right\}.$$

We define the maximum likelihood estimator $\widehat{\psi}_n := (\widehat{\beta}_n', \widehat{\theta}_n')'$ of ψ as the solution (if it exists) of the k-dimensional score equation

$$\dot{l}_n(\psi) = \frac{\partial l_n(\psi)}{\partial \psi} = 0, \tag{4}$$

where $l_n(\psi) := \log L_n(\psi)$ is the log-likelihood function. In the following, we shall be interested in the asymptotic properties of the maximum likelihood estimator $\widehat{\beta}_n$ of β, considered as a sub-component of $\widehat{\psi}_n$. We will however obtain consistency and asymptotic normality results for the whole $\widehat{\psi}_n$. Before proceeding, we need to set some further notations.

2.3. Some further notations

Define first the $(p \times n)$ and $(q \times n)$ matrices

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{pmatrix} \qquad \text{and} \qquad \mathbb{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ Z_{12} & Z_{22} & \cdots & Z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1q} & Z_{2q} & \cdots & Z_{nq} \end{pmatrix},$$

and let $\mathbb{W}$ be the $(k \times 2n)$ block-matrix defined as

$$\mathbb{W} = \begin{bmatrix} \mathbb{X} & 0_{pn} \\ 0_{qn} & \mathbb{Z} \end{bmatrix},$$

where 0_{ab} denotes the $(a \times b)$ matrix whose components are all equal to zero (for any positive integer values a, b). Let also $C(\psi)$ be the $2n$-dimensional column vector defined as

$$C(\psi) = \left((A^{\beta}(\psi) - B^{\beta}(\psi))', (A^{\theta}(\psi) - B^{\theta}(\psi))'\right)',$$

where $A^{\beta}(\psi) = (A_i^{\beta}(\psi))_{1 \le i \le n}$, $B^{\beta}(\psi) = (B_i^{\beta}(\psi))_{1 \le i \le n}$, $A^{\theta}(\psi) = (A_i^{\theta}(\psi))_{1 \le i \le n}$, and $B^{\theta}(\psi) = (B_i^{\theta}(\psi))_{1 \le i \le n}$ are n-dimensional column vectors with respective elements

$$A_i^{\beta}(\psi) = \frac{1 + e^{\theta' \mathbf{Z}_i}}{1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i}} Y_i, \quad B_i^{\beta}(\psi) = \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\beta' \mathbf{X}_i})(1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i})},$$
$$A_i^{\theta}(\psi) = \frac{1 + e^{\beta' \mathbf{X}_i}}{1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i}} Y_i, \quad B_i^{\theta}(\psi) = \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\theta' \mathbf{Z}_i})(1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i})}.$$

Then, simple algebra shows that the score equation can be rewritten as

$$\dot{l}_n(\psi) = \mathbb{W} C(\psi) = 0.$$

If $M = (M_{ij})_{1 \le i \le a, 1 \le j \le b}$ denotes some $(a \times b)$ matrix, we will denote by $M_{\bullet j}$ its j-th column $(j = 1, \ldots, b)$ that is, $M_{\bullet j} = (M_{1j}, \ldots, M_{aj})'$. Then, it will be useful to rewrite the score vector as

$$\dot{l}_n(\psi) = \sum_{j=1}^{2n} \mathbb{W}_{\bullet j} C_j(\psi).$$

We shall further note $\ddot{l}_n(\psi)$ the $(k \times k)$ matrix of second derivatives of $l_n(\psi)$ that is, $\ddot{l}_n(\psi) = \partial^2 l_n(\psi) / \partial\psi \partial\psi'$. Let $\mathbb{D}(\psi) = (\mathbb{D}_{ij}(\psi))_{1 \le i,j \le 2n}$ be the $(2n \times 2n)$ block matrix defined as

$$\mathbb{D}(\psi) = \begin{bmatrix} \mathbb{D}_1(\psi) & \mathbb{D}_3(\psi) \\ \mathbb{D}_3(\psi) & \mathbb{D}_2(\psi) \end{bmatrix},$$

where $\mathbb{D}_1(\psi), \mathbb{D}_2(\psi)$, and $\mathbb{D}_3(\psi)$ are $(n \times n)$ diagonal matrices, with i-th diagonal elements $(i = 1, \ldots, n)$ respectively given by

$$\mathbb{D}_{1,ii}(\psi) = \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\beta' \mathbf{X}_i})^2 (1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i})},$$
$$\mathbb{D}_{2,ii}(\psi) = \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\theta' \mathbf{Z}_i})^2 (1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i})},$$
$$\mathbb{D}_{3,ii}(\psi) = \frac{e^{\beta' \mathbf{X}_i + \theta' \mathbf{Z}_i}}{(1 + e^{\beta' \mathbf{X}_i})(1 + e^{\theta' \mathbf{Z}_i})(1 + e^{\beta' \mathbf{X}_i} + e^{\theta' \mathbf{Z}_i})}.$$

Then, some algebra shows that $\ddot{l}_n(\psi)$ can be expressed as

$$\ddot{l}_n(\psi) = -\mathbb{W}\mathbb{D}(\psi)\mathbb{W}'.$$

Note that the size of $C(\psi), \mathbb{W}$, and $\mathbb{D}(\psi)$ depends on n. However, in order to simplify notations, n will not be used as a lower index for these vector and matrices. In the next section, we investigate the question of parameter identifiability in model (1)-(2)-(3).

3. Identifiability and regularity conditions

We first state some regularity conditions that will be needed to ensure identifiability and the asymptotic results in Section 4:

C1 The covariates are bounded that is, there exist compact sets $F \subset \mathbb{R}^p$ and $G \subset \mathbb{R}^q$ such that $\mathbf{X}_i \in F$ and $\mathbf{Z}_i \in G$ for every $i = 1, 2, \ldots$ For every $i = 1, 2, \ldots$, $j = 2, \ldots, p$, $k = 2, \ldots, q$, $\text{var}[X_{ij}] > 0$ and $\text{var}[Z_{ik}] > 0$. For every $i = 1, 2, \ldots$, the X_{ij} $(j = 1, \ldots, p)$ are linearly independent, and the Z_{ik} $(k = 1, \ldots, q)$ are linearly independent.

C2 Let $\psi_0 = (\beta_0', \theta_0')'$ denote the true parameter value. β_0 and θ_0 lie in the interior of known compact sets $\mathcal{B} \subset \mathbb{R}^p$ and $\mathcal{G} \subset \mathbb{R}^q$ respectively.

C3 The Hessian matrix $\ddot{l}_n(\psi)$ is negative definite and of full rank, for every $n = 1, 2, \ldots$ Let λ_n and Λ_n be respectively the smallest and largest eigenvalues of $\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}'$. There exists a finite positive constant c_2 such that $\Lambda_n/\lambda_n < c_2$ for every $n = 1, 2, \ldots$

C4 There exists a continuous covariate V which is in $\mathbf{X}$ but not in $\mathbf{Z}$ that is, if β_V and θ_V denote the coefficients of V in the linear predictors (1) and (3) respectively, then $\beta_V \neq 0$ and $\theta_V = 0$. At a model-building stage, it is known that V is in $\mathbf{X}$.

The conditions C1, C2, C3 are classical conditions for identifiability and asymptotic results in standard logistic regression (see, for example, Gouriéroux and Monfort (1981) and Guyon (2001)). The condition C4, which imposes some restrictions on the covariates, is required for identifiability of ψ in the joint model (1)-(2)-(3) (we may alternatively assume that the continuous covariate V is in $\mathbf{Z}$ but not in $\mathbf{X}$). In the following, we will assume that V is in $\mathbf{X}$ but not in $\mathbf{Z}$, with $\beta_V := \beta_l$ for some $l \in \{2, \ldots, p\}$, and for the i-th individual, we will denote V_i by X_{il}. The condition C4 is discussed in greater details in the following two remarks.

Remark 2. We may relate the identifiability issue in model (1)-(2)-(3) to the problem of identifiability of mixtures of logistic regression models, which was investigated by Follmann and Lambert (1991). Follmann and Lambert (1991) considered the case where there is a finite number c of components in the mixture (we consider here the case where $c = 2$, with one degenerate component) and the mixing probabilities are constant (here, the mixing probabilities given by (3) are allowed to depend on covariates). The authors have shown that finite mixtures of logistic regressions are identifiable provided that the number of unique covariate combinations values is sufficiently large. C4 can be viewed as a sufficient condition for achieving the same kind of requirement. A similar condition appears in Kelley and Anderson (2008).
To understand C4, note that if $\mathbf{X}_i = \mathbf{Z}_i$, then exchanging the parameters β and θ in (1) and (3) yields the same likelihood value $L_n(\psi)$, which is a cause of model non-identifiability. A

similar remark holds if we invert the linear predictors $\beta'\mathbf{X}_i$ and $\theta'\mathbf{Z}_i$. The condition C4 evacuates these problems.
First, by asking one of the covariates to be significant in one and only one linear predictor, C4 prevents $\beta'\mathbf{X}$ and $\theta'\mathbf{Z}$ from being of the same form, and the parameters are thus not exchangeable. Secondly, by assuming that we know, prior to model fitting, that there exists a covariate V which is in $\mathbf{X}$ but not in $\mathbf{Z}$, C4 will force each linear predictor to be attached to the correct corresponding model (1) or (3).
These facts are illustrated in a web-based supplementary document available at the following address: `http://perso.univ-lr.fr/jfdupuy/supplementary.pdf`. There, we provide the results of a simulation study which investigates numerically the identifiability of model (1)-(2)-(3). For each of the models considered in this study, we assume that C4 is satisfied: the linear predictors $\beta'\mathbf{X}_i$ and $\theta'\mathbf{Z}_i$ share three covariates (one is continuous, two are discrete), and an additional continuous covariate is included in $\mathbf{X}_i$. Using the procedure described in Section 2, maximum likelihood estimates are obtained for β and θ, and are averaged over $N = 1000$ samples (we considered several combinations of sample size, proportion of immunes, proportion of infected among the susceptibles). Both parameters β and θ appear to be identifiable (the averaged estimates appear to be close to the true parameters, including those corresponding to the three shared covariates).

Remark 3. The condition C4 does not appear to be too restrictive in practice. Consider the example of the transmission of some disease by breastfeeding. If every child in the sample is breastfeeded, it can be expected that the length (in days, say) of the breastfeeding period (a continuous covariate) will influence the probability of infection, while the susceptibility probability will rather depend on risk factors such as say, the mother's infection status. It is also worth noting that the consequences of C4, in terms of model-building, are rather mild. At a model-building stage, we may be tempted to incorporate all available covariates in both linear predictors (1) and (3), and to remove irrelevant factors by using backward elimination. The condition C4 slightly restricts this fitting strategy, by imposing that one relevant continuous covariate is incorporated in one (and only one) linear predictor. This should often be doable in practice, since the statistician often gets some prior knowledge (from the clinicians, epidemiologists,...) about the dataset to be analyzed.

We are now in position to prove the following result:

Theorem 1 (Identifiability). *Under the conditions C1-C4, the model (1)-(2)-(3) is identifiable; that is, $L_1(\psi) = L_1(\psi^*)$ almost surely implies $\psi = \psi^*$.*

Proof of Theorem 1. Suppose that $L_1(\psi) = L_1(\psi^*)$ almost surely. Under C1 and C2, there exists a positive constant c_1 such that for every $\mathbf{x} \in F$, $\mathbf{z} \in G$, and $\psi \in \mathcal{B} \times \mathcal{G}$, $c_1 < \mathbb{P}(Y = 1|\mathbf{x}, \mathbf{z}) < 1 - c_1$. Thus we can find a $\omega \in \Omega$, outside the negligible set where $L_1(\psi) \neq L_1(\psi^*)$, and such that $Y(\omega) = 1$ when $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$. For this ω, $L_1(\psi) = L_1(\psi^*)$ becomes

$$\frac{e^{\beta'\mathbf{x}+\theta'\mathbf{z}}}{(1+e^{\beta'\mathbf{x}})(1+e^{\theta'\mathbf{z}})} = \frac{e^{\beta^{*'}\mathbf{x}+\theta^{*'}\mathbf{z}}}{(1+e^{\beta^{*'}\mathbf{x}})(1+e^{\theta^{*'}\mathbf{z}})}.$$

This can be rewritten as

$$\frac{1+e^{-\beta'\mathbf{x}}}{1+e^{-\beta^{*'}\mathbf{x}}} = \frac{1+e^{-\theta^{*'}\mathbf{z}}}{1+e^{-\theta'\mathbf{z}}}. \tag{5}$$

Now, under the condition C4, taking the partial derivative of both sides of (5) with respect to the l-th component of $\mathbf{x}$ (X_{il} is a continuous covariate) yields

$$\frac{-\beta_l e^{-\beta'\mathbf{x}}(1+e^{-\beta^{*'}\mathbf{x}}) + \beta_l^* e^{-\beta^{*'}\mathbf{x}}(1+e^{-\beta'\mathbf{x}})}{(1+e^{-\beta^{*'}\mathbf{x}})^2} = 0$$

since the right-hand-side of (5) does not depend on $\mathbf{x}$. Thus, it follows that

$$\frac{\beta_l}{\beta_l^*} = \frac{1+e^{\beta'\mathbf{x}}}{1+e^{\beta^{*'}\mathbf{x}}}.$$

Differentiating both sides of this equality with respect to the l-th component of $\mathbf{x}$ further yields $(\beta-\beta^*)'\mathbf{x} = 0$, which implies that $\beta = \beta^*$ under C1. It remains to show that $\theta = \theta^*$, which reduces to the identifiability problem in the standard logistic regression model. We have that $\theta = \theta^*$ under C1 (see Guyon (2001) for example), which concludes the proof.

□

We now turn to the asymptotic theory for the proposed estimator.

4. Asymptotic theory

In this section, we establish rigorously the existence, consistency and asymptotic normality of the maximum likelihood estimator $\widehat{\beta}_n$ of β in model (1), obtained from a sample of binary response data with a cure fraction. In the sequel, the space $\mathbb{R}^k$ of k-dimensional (column) vectors will be provided with the Euclidean norm, and the space $\mathbb{R}^{k\times k}$ of $(k\times k)$ real matrices will be provided with the spectral norm (we will use the same notation $\|\cdot\|$ for both). We first prove the following result:

Theorem 2 (Existence and consistency). *Under the conditions C1-C3, the maximum likelihood estimator $\widehat{\psi}_n$ exists almost surely as $n\to\infty$, and converges almost surely to ψ_0, if and only if λ_n tends to infinity as $n\to\infty$.*

Proof of Theorem 2. The principle of the proof is similar to Gouriéroux and Monfort (1981) but the technical details are different. Three lemmas are needed. The first lemma essentially provides an intermediate technical result. Its proof is postponed to the appendix.

Lemma 1. *Let $\phi_n : \mathbb{R}^k \longrightarrow \mathbb{R}^k$ be defined as*

$$\phi_n(\psi) = \psi + (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi).$$

Then there exists an open ball $B(\psi_0, r)$ (with $r>0$) such that ϕ_n satisfies the Lipschitz condition on $B(\psi_0, r)$ that is,

$$\|\phi_n(\psi_1) - \phi_n(\psi_2)\| \le c\,\|\psi_1-\psi_2\| \ \text{ for all } \psi_1,\psi_2 \in B(\psi_0,r), \tag{6}$$

and $0<c<1$.

Lemma 2. *The maximum likelihood estimator $\widehat{\psi}_n$ exists almost surely as $n\to\infty$, and converges almost surely to ψ_0, if and only if $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi_0)$ converges almost surely to 0.*

Proof of Lemma 2. We first prove that the condition is sufficient. Thus, we assume that $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi_0)$ converges almost surely to 0.
Define $\eta_n(\psi) = \psi - \phi_n(\psi) = -(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi)$ and let ϵ be an arbitrary positive value. Then for almost every $\omega \in \Omega$, there exists an integer value $n(\epsilon, \omega)$ such that for any $n \geq n(\epsilon, \omega)$, $\|\eta_n(\psi_0)\| \leq \epsilon$ or equivalently, $0 \in B(\eta_n(\psi_0), \epsilon)$. In particular, let $\epsilon = (1-c)s$ with $0 < c < 1$ such as in Lemma 1. Since ϕ_n satisfies the Lipschitz condition (6) (by Lemma 1), the lemma 2 of Gouriéroux and Monfort (1981) ensures that there exists an element of $B(\psi_0, s)$ (let denote this element by $\widehat{\psi}_n$) such that $\eta_n(\widehat{\psi}_n) = 0$ that is,

$$(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\widehat{\psi}_n) = 0.$$

The condition C3 implies that $\dot{l}_n(\widehat{\psi}_n) = 0$ and that $\widehat{\psi}_n$ is the unique maximizer of l_n. To summarize, we have shown that for almost every $\omega \in \Omega$ and for every $s > 0$, there exists an integer value $n(s, \omega)$ such that if $n \geq n(s, \omega)$, then the maximum likelihood estimator $\widehat{\psi}_n$ exists, and $\left\|\widehat{\psi}_n - \psi_0\right\| \leq s$ (that is, $\widehat{\psi}_n$ converges almost surely to ψ_0).
We now prove that the condition that $\eta_n(\psi_0)$ converges almost surely to 0 is necessary. We use a proof by contradiction.
Assume that as $n \to \infty$, $\widehat{\psi}_n$ exists and converges almost surely to ψ_0, but $\eta_n(\psi_0)$ does not converge almost surely to 0. Then there exists a set $\widetilde{\Omega} \subset \Omega$ with $\mathbb{P}(\widetilde{\Omega}) > 0$, such that if $\omega \in \widetilde{\Omega}$, there exists $\epsilon > 0$ such that for every $m \in \mathbb{N}$, there exists $n \geq m$ with $\|\eta_n(\psi_0)\| > \epsilon$. Now, let $t = \frac{\epsilon}{d(1+c)}$, with $d > 1$ sufficiently large so that $t \leq r$, where r is such as in Lemma 1. Then for every $\psi \in B(\psi_0, t)$, the following holds:

$$\begin{aligned} \|\eta_n(\psi_0) - \eta_n(\psi)\| &= \|\psi_0 - \phi_n(\psi_0) - \psi + \phi_n(\psi)\| \\ &\leq \|\psi_0 - \psi\| + \|\phi_n(\psi) - \phi_n(\psi_0)\| \\ &\leq t(1+c) = \frac{\epsilon}{d}, \end{aligned}$$

where the second to third line follows by Lemma 1. Therefore, for every $\psi \in B(\psi_0, t)$,

$$\epsilon < \|\eta_n(\psi_0)\| \leq \|\eta_n(\psi_0) - \eta_n(\psi)\| + \|\eta_n(\psi)\| \leq \|\eta_n(\psi)\| + \frac{\epsilon}{d}$$

and we conclude that for every $\psi \in B(\psi_0, t)$, $\|\eta_n(\psi)\| > \epsilon(1 - \frac{1}{d}) > 0$. Since $\eta_n(\widehat{\psi}_n) = 0$, $\widehat{\psi}_n$ cannot belong to $B(\psi_0, t)$ for large n, which implies that $\widehat{\psi}_n$ does not converge almost surely to ψ_0. This is the desired contradiction.

□

Lemma 3. $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi_0)$ *converges almost surely to* 0 *if and only if* λ_n *tends to infinity as* $n \to \infty$.

Proof of Lemma 3. We first prove that the condition is sufficient that is, we assume that λ_n tends to infinity as $n \to \infty$. Define the $(2n \times k)$ matrix $\mathbb{V} = (\mathbb{D}(\psi_0))^{\frac{1}{2}}\mathbb{W}'$ and the $2n$-dimensional vector $U = (\mathbb{D}(\psi_0))^{-\frac{1}{2}}C(\psi_0)$. Then

$$\mathbb{E}[U] = 0 \text{ and } \text{var}[U] = I_{2n}, \tag{7}$$

where I_{2n} denotes the identity matrix of order $2n$. To see this, note that

$$\begin{aligned}\mathbb{E}[U] &= \mathbb{E}[\mathbb{E}[(\mathbb{D}(\psi_0))^{-\frac{1}{2}}C(\psi_0)|\mathbb{X},\mathbb{Z}]] \\ &= \mathbb{E}[(\mathbb{D}(\psi_0))^{-\frac{1}{2}}\mathbb{E}[C(\psi_0)|\mathbb{X},\mathbb{Z}]] \\ &= \mathbb{E}[(\mathbb{D}(\psi_0))^{-\frac{1}{2}}\mathbb{E}[\big((A^{\beta}(\psi_0)-B^{\beta}(\psi_0))',(A^{\theta}(\psi_0)-B^{\theta}(\psi_0))'\big)'|\mathbb{X},\mathbb{Z}]].\end{aligned}$$

For every $i=1,\ldots,n$, $\mathbb{E}[A_i^{\beta}(\psi_0)-B_i^{\beta}(\psi_0)|\mathbb{X},\mathbb{Z}]] = \mathbb{E}[A_i^{\beta}(\psi_0)-B_i^{\beta}(\psi_0)|\mathbf{X}_i,\mathbf{Z}_i]]$ by independence between the individuals, and

$$\begin{aligned}\mathbb{E}[A_i^{\beta}(\psi_0)-B_i^{\beta}(\psi_0)|\mathbf{X}_i,\mathbf{Z}_i]] &= \frac{1+e^{\theta_0'\mathbf{Z}_i}}{1+e^{\beta_0'\mathbf{X}_i}+e^{\theta_0'\mathbf{Z}_i}}\mathbb{P}(Y_i=1|\mathbf{X}_i,\mathbf{Z}_i)-B_i^{\beta}(\psi_0) \\ &= B_i^{\beta}(\psi_0)-B_i^{\beta}(\psi_0) \\ &= 0.\end{aligned}$$

Similarly, $\mathbb{E}[A_i^{\theta}(\psi_0)-B_i^{\theta}(\psi_0)|\mathbb{X},\mathbb{Z}]]=0$ for every $i=1,\ldots,n$ and thus, $\mathbb{E}[C(\psi_0)|\mathbb{X},\mathbb{Z}]]=0$ and $\mathbb{E}[U]=0$.
Next, $\text{var}[U]=\mathbb{E}[\text{var}[U|\mathbb{X},\mathbb{Z}]]$ since $\mathbb{E}[U|\mathbb{X},\mathbb{Z}]=0$. Moreover,

$$\text{var}[U|\mathbb{X},\mathbb{Z}] = (\mathbb{D}(\psi_0))^{-\frac{1}{2}}\text{var}[C(\psi_0)|\mathbb{X},\mathbb{Z}](\mathbb{D}(\psi_0))^{-\frac{1}{2}},$$

with $\text{var}[C(\psi_0)|\mathbb{X},\mathbb{Z}] = \text{var}[\big(A^{\beta}(\psi_0)',A^{\theta}(\psi_0)'\big)'|\mathbb{X},\mathbb{Z}]$ a $(2n\times 2n)$ block-matrix of the form

$$\begin{bmatrix}\mathbb{V}_1 & \mathbb{V}_3\\ \mathbb{V}_3 & \mathbb{V}_2\end{bmatrix}$$

where $\mathbb{V}_1,\mathbb{V}_2$, and $\mathbb{V}_3$ are $(n\times n)$ matrices. The i-th diagonal elements $(i=1,\ldots,n)$ of $\mathbb{V}_1,\mathbb{V}_2$, and $\mathbb{V}_3$ are $\text{var}[A_i^{\beta}(\psi_0)|\mathbb{X},\mathbb{Z}]$, $\text{var}[A_i^{\theta}(\psi_0)|\mathbb{X},\mathbb{Z}]$, and $\text{cov}[A_i^{\beta}(\psi_0),A_i^{\theta}(\psi_0)|\mathbb{X},\mathbb{Z}]$ respectively. Similar calculations as above yield: $\text{var}[A_i^{\beta}(\psi_0)|\mathbb{X},\mathbb{Z}]=\mathbb{D}_{1,ii}(\psi_0)$, $\text{var}[A_i^{\theta}(\psi_0)|\mathbb{X},\mathbb{Z}]=\mathbb{D}_{2,ii}(\psi_0)$, and $\text{cov}[A_i^{\beta}(\psi_0),A_i^{\theta}(\psi_0)|\mathbb{X},\mathbb{Z}]=\mathbb{D}_{3,ii}(\psi_0)$. Note also that $\mathbb{V}_1,\mathbb{V}_2$, and $\mathbb{V}_3$ are diagonal matrices, by independence between the individuals. It follows that $\text{var}[C(\psi_0)|\mathbb{X},\mathbb{Z}]=\mathbb{D}(\psi_0)$ and thus, $\text{var}[U|\mathbb{X},\mathbb{Z}]=I_{2n}$ and $\text{var}[U]=I_{2n}$.
By Gouriéroux and Monfort (1981) (proof of Lemma 4), if (7) holds, $\Lambda_n/\lambda_n<c_2$ for every $n=1,2,\ldots$, and λ_n tends to infinity as $n\to\infty$, then

$$(\mathbb{V}'\mathbb{V})^{-1}\mathbb{V}'U \xrightarrow{a.s.} 0 \text{ as } n\to\infty$$

that is, $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi_0)$ converges almost surely to 0.
We now prove that the condition is necessary. Assume that λ_n does not tend to infinity as $n\to\infty$. By Gouriéroux and Monfort (1981) (proof of Lemma 4), $(\mathbb{V}'\mathbb{V})^{-1}\mathbb{V}'U$ (and therefore $(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\dot{l}_n(\psi_0)$) cannot converge to 0, which concludes the proof.

□

Finally, Theorem 2 follows by Lemma 2 and Lemma 3.

□

We now turn to the convergence in distribution of the proposed estimator, which is stated by the following theorem:

Theorem 3 (Asymptotic normality). *Assume that the conditions C1-C3 hold and that* $\widehat{\psi}_n$ *converges almost surely to* ψ_0*. Let* $\widehat{\Sigma}_n = \mathbb{W}\mathbb{D}(\widehat{\psi}_n)\mathbb{W}'$ *and* I_k *denote the identity matrix of order* k*. Then* $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ *converges in distribution to the Gaussian vector* $\mathcal{N}(0, I_k)$*.*

Proof of Theorem 3. A Taylor expansion of the score function is as

$$0 = \dot{l}_n(\widehat{\psi}_n) = \dot{l}_n(\psi_0) + \ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$$

where $\widetilde{\psi}_n$ lies between $\widehat{\psi}_n$ and ψ_0, and thus $\dot{l}_n(\psi_0) = -\ddot{l}_n(\widetilde{\psi}_n)(\widehat{\psi}_n - \psi_0)$. Let $\widetilde{\Sigma}_n := -\ddot{l}_n(\widetilde{\psi}_n) = \mathbb{W}\mathbb{D}(\widetilde{\psi}_n)\mathbb{W}'$ and $\Sigma_{n,0} := \mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}'$. Now,

$$\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0) = \left[\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}}\right]\left[\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}}\right]\Sigma_{n,0}^{-\frac{1}{2}}\left(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0)\right). \tag{8}$$

The two terms in brackets in (8) converge almost surely to I_k. To see this, we show for example that $\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \xrightarrow{a.s.} 0$ as $n \to \infty$. First, note that

$$\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\| \leq \Lambda_n^{\frac{1}{2}}\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\|\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|, \tag{9}$$

and

$$\Lambda_n^{-1}\left\|\Sigma_{n,0} - \widetilde{\Sigma}_n\right\| = \Lambda_n^{-1}\left\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{W}'\right\|.$$

Note also that $\widetilde{\psi}_n$ converges almost surely to ψ_0 (that is, for every $\omega \in \breve{\Omega}$, where $\breve{\Omega} \subset \Omega$ and $\mathbb{P}(\breve{\Omega}) = 1$). Let $\omega \in \breve{\Omega}$. By the same arguments as in the proof of Lemma 1, for every $\epsilon > 0$, there exists a positive $n(\epsilon, \omega) \in \mathbb{N}$ such that if $n \geq n(\epsilon, \omega)$, then $\Lambda_n^{-1}\left\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{W}'\right\| \leq \epsilon$. Hence $\Lambda_n^{-1}\left\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\widetilde{\psi}_n))\mathbb{W}'\right\|$ converges almost surely to 0. By continuity of the map $x \mapsto x^{\frac{1}{2}}$, $\left\|\Lambda_n^{-\frac{1}{2}}\left(\Sigma_{n,0}^{\frac{1}{2}} - \widetilde{\Sigma}_n^{\frac{1}{2}}\right)\right\|$ converges also almost surely to 0. Moreover, for n sufficiently large, there exists a positive constant $c_4 < \infty$ such that almost surely, $\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\right\| \leq c_4\lambda_n^{-\frac{1}{2}}$. It follows from (9) and the condition C3 that $\left\|\widetilde{\Sigma}_n^{-\frac{1}{2}}\Sigma_{n,0}^{\frac{1}{2}} - I_k\right\|$ converges almost surely to 0. The almost sure convergence to 0 of $\left\|\widehat{\Sigma}_n^{\frac{1}{2}}\widetilde{\Sigma}_n^{-\frac{1}{2}} - I_k\right\|$ follows by similar arguments.

It remains for us to show that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$, or equivalently, that $(\mathbb{V}'\mathbb{V})^{-\frac{1}{2}}\mathbb{V}'U$ converges in distribution to $\mathcal{N}(0, I_k)$. Following Eicker (1966), this convergence holds if we can check the following conditions: i) $\max_{1\leq i\leq 2n}\mathbb{V}_{i\bullet}\Sigma_{n,0}^{-1}\mathbb{V}_{i\bullet}' \longrightarrow 0$ as $n \to \infty$, ii) $\sup_{1\leq i\leq 2n}\mathbb{E}[U_i^2 1_{\{|U_i|>\alpha\}}] \longrightarrow 0$ as $\alpha \to \infty$, iii) $\inf_{1\leq i\leq 2n}\mathbb{E}[U_i^2] > 0$, where $\mathbb{V}_{i\bullet}$ and U_i respectively denote the i-th raw of $\mathbb{V}$ and the i-th component of U, $i = 1, \ldots, 2n$. Condition i) follows by noting that

$$0 \leq \max_{1\leq i\leq 2n}\mathbb{V}_{i\bullet}\Sigma_{n,0}^{-1}\mathbb{V}_{i\bullet}' \leq \max_{1\leq i\leq 2n}\|\mathbb{V}_{i\bullet}\|^2\left\|\Sigma_{n,0}^{-1}\right\| = \max_{1\leq i\leq 2n}\frac{1}{\lambda_n}\|\mathbb{V}_{i\bullet}\|^2,$$

and that $\|\mathbb{V}_{i\bullet}\|$ is bounded above, by C1 and C2. Moreover, $\frac{1}{\lambda_n}$ tends to 0 as $n \to \infty$, since $\widehat{\psi}_n$ converges almost surely to ψ_0. Condition ii) follows by noting that the components U_i of U are bounded under C1 and C2. Finally, for every $i = 1, \ldots, 2n$, $\mathbb{E}[U_i^2] = \mathrm{var}[U_i]$ since U

is centered. We have proved (see Lemma 3) that $\text{var}[U] = I_{2n}$, thus for every $i = 1, \ldots, 2n$, $\text{var}[U_i] = 1$, and finally, $\inf_{1 \leq i \leq 2n} \mathbb{E}[U_i^2] = 1 > 0$.

To summarize, we have proved that $\Sigma_{n,0}^{-\frac{1}{2}}(\widetilde{\Sigma}_n(\widehat{\psi}_n - \psi_0))$ converges in distribution to $\mathcal{N}(0, I_k)$.

This result, combined with Slutsky's theorem and equation (8), implies that $\widehat{\Sigma}_n^{\frac{1}{2}}(\widehat{\psi}_n - \psi_0)$ converges in distribution to $\mathcal{N}(0, I_k)$.

□

5. A simulation study and real data example

5.1. Study design

In this section, we investigate the numerical properties of the maximum likelihood estimator $\widehat{\beta}_n$, under various conditions. The simulation setting is as follows. We consider the following models for the infection status:

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta_1 + \beta_2 X_{i2} & \text{if } S_i = 1 \\ \mathbb{P}(Y = 1|\mathbf{X}_i, S_i) = 0 & \text{if } S_i = 0 \end{cases}$$

and the immunity status:

$$\log\left(\frac{\mathbb{P}(S = 1|\mathbf{Z}_i)}{1 - \mathbb{P}(S = 1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2},$$

where X_{i2} is normally distributed with mean 0 and variance 1, and Z_{i2} is normally distributed with mean 1 and variance 1. An i.i.d. sample of size n of the vector $(Y, S, \mathbf{X}, \mathbf{Z})$ is generated from this model, and for each individual i, we get a realization $(y_i, s_i, \mathbf{x}_i, \mathbf{z}_i)$, where s_i is considered as unknown if $y_i = 0$. A maximum likelihood estimator $\widehat{\beta}_n$ of $\beta = (\beta_1, \beta_2)$ is obtained from this incomplete dataset by solving the score equation (4), using the `optim` function of the software `R`. An estimate is also obtained for $\theta = (\theta_1, \theta_2)$, but θ is not the primary parameter of interest hence we only focus on the simulation results for $\widehat{\beta}_n$.

The finite-sample behavior of the maximum likelihood estimator $\widehat{\beta}_n$ was assessed for several sample sizes ($n = 100, 500, 1000, 1500$) and various values for the percentage of immunes in the sample, namely 25%, 50%, and 75%. The case where it is known that there are no immunes in the sample was also considered. In this case, there is no missing information about the infection status and therefore, this case provides a benchmark for evaluating the performance of the proposed estimation method. We also considered different values for the proportion of infected individuals among the susceptibles. The desired proportions of immunes and infected were obtained by choosing appropriate values for the parameters β (the parameter of interest) and θ (the nuisance parameter). The following values were considered for β: i) model $\mathcal{M}_1$: $\beta = (-.8, 1)$ (using these values, approximately 30% of the susceptibles are infected), ii) model $\mathcal{M}_2$: $\beta = (1, .7)$ (approximately 70% of the susceptibles are infected), iii) model $\mathcal{M}_3$: $\beta = (-.8, 0)$ (approximately 30% of the susceptibles are infected), iv) model $\mathcal{M}_4$: $\beta = (1, 0)$ (approximately 70% of the susceptibles are infected).

5.2. Results

For each configuration (sample size, percentage of immunes, percentage of infected among susceptibles) of the design parameters, $N = 1500$ samples were obtained. Based on these

1500 repetitions, we obtain averaged values for the estimates of β_1 and β_2, which are calculated as $N^{-1}\sum_{j=1}^{N}\widehat{\beta}_{1,n}^{(j)}$ and $N^{-1}\sum_{j=1}^{N}\widehat{\beta}_{2,n}^{(j)}$, where $\widehat{\beta}_n^{(j)} = (\widehat{\beta}_{1,n}^{(j)}, \widehat{\beta}_{2,n}^{(j)})$ is the estimate obtained from the j-th simulated sample. For each of the parameters β_1 and β_2, we also obtain the empirical root mean square and mean absolute errors, based on the N samples. When $\beta_2 \neq 0$ (respectively $\beta_2 = 0$), we obtain the empirical power (respectively the empirical size) of the Wald test at the 5% level for testing $H_0 : \beta_2 = 0$ (models $\mathcal{M}_1$ and $\mathcal{M}_2$, see Tables 1 and 2) (respectively models $\mathcal{M}_3$ and $\mathcal{M}_4$, see Tables 1 and 2). The null hypothesis $H_0 : \beta_2 = 0$ is the hypothesis that the predictor X_2 does not influence the risk of infection of susceptible individuals. The results are summarized in Tables 1 and 2.

Tables 1 to 2 about here

From these tables, it appears that the proposed maximum likelihood estimator $\widehat{\beta}_n$ provides a reasonable approximation of the true parameter value, even when the percentage of immunes is high. While the bias of $\widehat{\beta}_n$ stays limited, its variability increases with the immune fraction, sometimes drastically when the sample size is small. Consequently, when the sample size is small ($n = 100$) and/or the immune proportion is very high (75%), the power of the Wald test for nullity of the regression coefficient β_2 can be low, compared to the case where there are no immunes. But we note that for moderately large to large sample sizes ($n \geq 500$), the dispersion indicators and the power of the Wald test indicate good performance of the maximum likelihood estimate, even when the immune proportion is up to 50%. The level of the Wald test for nullity of β_2 is globally respected except, for every immune proportion, when the sample size is small ($n = 100$).

We compare these results to the ones obtained from a "naive" method where: i) we consider every individual i such that $\{Y_i = 0\}$ as being susceptible but uninfected, that is we ignore the eventual immunity of this individual, ii) we apply a usual logistic regression analysis to the resulting dataset. The results of such "naive" analysis for model $\mathcal{M}_1$ are given in Table 3 (the results for models $\mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ yield similar observations and thus, they are not given here. However, the complete simulation study is available from the web-based supplementary document mentioned above).

Table 3 about here

From this table, it appears that ignoring the immunity present in the sample results in strongly biased estimates of β. The bias of the intercept estimate increases with the immune proportion. At the same time, the estimate of the regression coefficient β_2 is biased towards 0 for all values of the immune percentage and sample size. This results in a very low power for the Wald test of nullity of β_2, and in a wrong interpretation of the relationship between the covariate X_2 and the binary response Y.

The quality of the Gaussian approximation to the large-sample distribution of $\widehat{\beta}_{2,n}$ was also investigated. For each configuration of the design parameters, histograms of the $\widehat{\beta}_{2,n}^{(j)}$ ($j = 1, \ldots, N$) are obtained, along with the corresponding QQ-plots. The plots for the model $\mathcal{M}_1$ are pictured on Figures 1 to 4 (the plots for the models $\mathcal{M}_2$, $\mathcal{M}_3$, $\mathcal{M}_4$ are given in the web-based file).

Figures 1 to 4 about here

From these figures, it appears that the normal approximation stated in Theorem 3 is reasonably satisfied when the proportion of immunes is moderate (25%), provided that the sample size is sufficiently large ($n \geq 500$, say). Consider the case when $\beta_2 \neq 0$. When the immune

fraction is large (50%), the normal approximation still appears reasonable, provided that the sample size is at least 1000, or eventually 1500. When the immune proportion is very large (75%), the distribution of $\widehat{\beta}_{2,n}$ can be highly skewed, in particular when the sample size is small. Consider the case when $\beta_2 = 0$. Then the finite-sample distribution of $\widehat{\beta}_{2,n}$ appears to be symmetric, with heavy tails however, especially when the sample size is small. When the immune fraction is about 50% and the sample size is greater than or equal to 500, the normal distribution appears to fit reasonably well the distribution of $\widehat{\beta}_{2,n}$.
Overall, these results indicate that a reliable statistical inference on the regression effect in the model (1) with a cure fraction should be based on a sample having, at least, a moderately large size ($n \geq 500$, say) when the immune fraction is moderate (25%), or a large size ($n \geq 1000$, say) when the immune proportion is large (50%).

5.3. A real data example

In this application, we consider a study of dengue fever, which is a mosquito-borne viral human disease. A dengue infection confers a partial and transient immunity against a subsequent infection (see Dussart et al. (2011)). We consider here a database of size $n = 528$, which was constituted with individuals recruited in Cambodia, Vietnam, French Guiana, and Brazil (Dussart et al. (2011)). Each individual i was diagnosed for dengue infection and coded as $Y_i = 1$ if infection was present and 0 otherwise. Note that if $Y_i = 0$, then the i-th individual may either be immune at the time of analysis (due to a temporary immunity acquired following a previous infection) or susceptible to dengue infection, albeit not infected. We aim at estimating the risk of infection for those individuals, based on this data set which also includes the following covariates: age (a continuous bounded covariate) and weight (coded 0 in case of underweight and 1 otherwise). We first ran a standard logistic regression analysis of the model logit $\mathbb{P}(Y = 1|\text{age, weight}) = \beta_1 + \beta_2\text{age} + \beta_3\text{weight}$. The results are displayed in Table 4. Then, we estimated the parameters β_1, β_2 and β_3 using the methodology described in Section 2.
Note first that the eventual immunity imparted by a past infection is only transient, thus there is no reason why an older individual (who has therefore been exposed longer to the risk of dengue fever) would have a greater probability of being immune than a younger one. In fact, individual susceptibility to the dengue infection may rather depend on whether the individual benefits or not from some preventive and control measures (such as the application of insecticides to larval habitats in his area, or appropriate water storage and waste disposal practices). Such informations are not available in our dataset.
Age was therefore taken as the variable V in condition C4, and we fitted to the data the model (1)-(2)-(3) with logit $\mathbb{P}(Y = 1|\text{age, weight}, S = 1) = \beta_1 + \beta_2\text{age} + \beta_3\text{weight}$ and logit $\mathbb{P}(S = 1|\text{ weight}) = \theta_1 + \theta_2\text{weight}$. Since the Wald-type test of "$\theta_2 = 0$" was not significant, we removed the weight from the model for susceptibility, resulting in a constant proportion of immunes. The final results of this fitting procedure are given in Table 4.

Table 4 about here

The fitted model produced the following estimate: $1-\exp(0.497)/(1+\exp(0.497)) \approx 0.38$ for the probability of being immune. Then, as expected, the estimated probabilities of infection obtained from our approach are larger than the ones derived from a standard analysis that does not take account of the possible immunity. For example, the probabilities of infection for individuals respectively aged 30 and 10 years old, with "normal" weight, are estimated by 0.29 and 0.55 (standard logistic regression) and 0.31 and 0.86 (from our approach). It is

expected that underweighted subjects (those considered to be under a healthy weight) will have higher risks of infection. The probabilities of infection for underweighted individuals respectively aged 30 and 10 years old are estimated by 0.48 and 0.73 (standard logistic regression) and 0.97 and 0.99 (our approach). While both approaches provide the same qualitative conclusions: the probability of dengue infection is higher for younger individuals and in case of underweight (caused by malnutrition for example), they differ on their estimations of the risk of infection. Our approach takes account of the eventual immunity imparted by a past infection and therefore, it is reasonable to think that the resulting estimations of the infection probabilities provide a more realistic picture of the infection risk for this data set. In particular, the estimates provided by our approach suggest that underweight constitutes a major risk factor for dengue infection, irrespectively of age.

6. Discussion and perspectives

In this paper, we have considered the problem of estimating the logistic regression model from a sample of binary response data with a cure fraction. The estimator we propose is obtained by maximizing a likelihood function, which is derived from a joint regression model for the binary response of interest and the cure indicator, considered as a random variable whose distribution is modeled by a logistic regression (the proposed joint model can thus be viewed as a zero-inflated Bernoulli regression model, with logit links for both the binary response of interest and the zero-inflation component). We have established the existence, consistency, and asymptotic normality of this estimator, and we have investigated its finite-sample properties via simulations.

Several open questions now deserve attention. The estimation approach proposed here relies on our ability to correctly specify the model for the binary immunity status. It is therefore of interest to investigate the effect of a misspecification of this model (and in particular, of the link function). The techniques and results by Czado and Santner (1992) may be useful for that purpose. Another issue of interest deals with the inference in the logistic regression model with a cure fraction, in a high-dimensional setting. We have established the theoretical properties of our estimator in a low-dimensional setting that is, when a small number of potential predictors are involved. Several recent contributions (see for example Huang et al. (2008) and Meier et al. (2008)) have considered the problem of estimation in the logistic model (without cure fraction) when the predictor dimension is much larger than the sample size (this problem arises, for example, in genetic studies where high-dimensional data are generated using microarray technologies). Extending our methodology to this setting constitutes another topic for further research.

Appendix

Proof of Lemma 1. Recall that I_k denotes the identity matrix of order k. Then we write:

$$\begin{aligned}
\left\| \frac{\partial \phi_n(\psi)}{\partial \psi'} \right\| &= \left\| I_k - (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\mathbb{W}\mathbb{D}(\psi)\mathbb{W}' \right\| \\
&= \left\| (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1}\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}' \right\| \\
&\leq \left\| (\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}')^{-1} \right\| \left\| \mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}' \right\| \\
&= \lambda_n^{-1} \left\| \mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}' \right\| .
\end{aligned}$$

Next, define $\mathcal{S} = \{(i,j) \in \{1,2,\ldots,2n\}^2 | \mathbb{D}_{ij}(\psi_0) \neq 0\}$. Then the following holds:

$$\begin{aligned}\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}'\| &= \left\| \sum_{i=1}^{2n}\sum_{j=1}^{2n} \mathbb{W}_{\bullet i}\mathbb{W}'_{\bullet j}(\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)) \right\| \\ &\leq \sum_{(i,j)\in\mathcal{S}} \left\|\mathbb{W}_{\bullet i}\mathbb{W}'_{\bullet j}\mathbb{D}_{ij}(\psi_0)\right\| \left| \frac{\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)}{\mathbb{D}_{ij}(\psi_0)} \right|.\end{aligned}$$

From C1 and C2, there exists a real constant c_3 ($c_3 > 0$) such that $\mathbb{D}_{ij}(\psi_0) > c_3$ for every $(i,j) \in \mathcal{S}$. Moreover, $\mathbb{D}_{ij}(\cdot)$ is uniformly continuous on $\mathcal{B} \times \mathcal{G}$, thus for every $\epsilon > 0$, there exists a positive r such that for all $\psi \in B(\psi_0, r)$, $|\mathbb{D}_{ij}(\psi) - \mathbb{D}_{ij}(\psi_0)| < \epsilon$. It follows that

$$\begin{aligned}\|\mathbb{W}(\mathbb{D}(\psi_0) - \mathbb{D}(\psi))\mathbb{W}'\| &\leq \frac{\epsilon}{c_3} \sum_{(i,j)\in\mathcal{S}} \left\|\mathbb{W}_{\bullet i}\mathbb{W}'_{\bullet j}\mathbb{D}_{ij}(\psi_0)\right\| \\ &\leq \frac{\epsilon}{c_3}\mathrm{tr}\left(\sum_{(i,j)\in\mathcal{S}} \mathbb{W}_{\bullet i}\mathbb{W}'_{\bullet j}\mathbb{D}_{ij}(\psi_0) \right) \\ &= \frac{\epsilon}{c_3}\mathrm{tr}\left(\sum_{i=1}^{2n}\sum_{j=1}^{2n} \mathbb{W}_{\bullet i}\mathbb{W}'_{\bullet j}\mathbb{D}_{ij}(\psi_0) \right) \\ &= \frac{\epsilon}{c_3}\mathrm{tr}\left(\mathbb{W}\mathbb{D}(\psi_0)\mathbb{W}'\right) \\ &\leq \frac{\epsilon}{c_3}\Lambda_n k.\end{aligned}$$

This in turn implies that $\left\| \frac{\partial \phi_n(\psi)}{\partial \psi'} \right\| \leq \frac{\epsilon \Lambda_n k}{c_3 \lambda_n} < \frac{\epsilon c_2 k}{c_3}$. Now, choosing $\epsilon = c\frac{c_3}{c_2 k}$ with $0 < c < 1$, we get that $\left\| \frac{\partial \phi_n(\psi)}{\partial \psi'} \right\| \leq c$ for all $\psi \in B(\psi_0, r)$, and the result follows.

□

References

Czado, C. and T. J. Santner (1992). The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference 33*(2), 213–231.

Dietz, E. and D. Böhning (2000). On estimation of the poisson parameter in zero-modified poisson models. *Computational Statistics & Data Analysis 34*, 441–459.

Dussart, P., L. Baril, L. Petit, L. Beniguel, L. C. Quang, S. Ly, R. do Socorro Azevedo, J.-B. Meynard, S. Vong, L. Chartier, A. Diop, O. Sivuth, V. Duong, C. M. Thang, M. Jacobs, A. Sakuntabhai, M. R. Texeira Nunes, V. T. Que Huong, P. Buchy, and P. F. Vasconcelos (2011). Study of dengue cases and the members of their households: a familial cluster analysis in the multinational denframe project. pp. Submitted.

Eicker, F. (1966). A multivariate central limit theorem for random linear vector forms. *Annals of Mathematical Statistics 37*, 1825–1828.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics 13*(1), 342–368.

Famoye, F. and K. P. Singh (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science 4*(1), 117–130.

Fang, H.-B., G. Li, and J. Sun (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics 32*(1), 59–75.

Follmann, D. A. and D. Lambert (1991). Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference 27*(3), 375–381.

Gouriéroux, C. and A. Monfort (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics 17*(1), 83–97.

Guyon, X. (2001). *Statistique et économétrie - Du modèle linéaire aux modèles non-linéaires*. Ellipses Marketing.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics 56*(4), 1030–1039.

Hilbe, J. M. (2009). *Logistic regression models*. Chapman & Hall: Boca Raton.

Hosmer, D. and S. Lemeshow (2000). *Applied logistic regression*. Wiley: New York.

Huang, J., S. Ma, and Z. C. H. (2008). The iterated lasso for high-dimensional logistic regression. *Technical report No. 392, The University of Iowa*.

Kelley, M. E. and S. J. Anderson (2008). Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. *Statistics in Medicine 27*(18), 3674–3688.

Lam, K. F., H. Xue, and Y. B. Cheung (2006). Semiparametric analysis of zero-inflated count data. *Biometrics 62*(4), 996–1003.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics 34*(1), 1–14.

Lee, A. H., K. Wang, J. A. Scott, K. K. W. Yau, and G. J. McLachlan (2006). Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research 15*, 47–61.

Lu, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics 60*(3), 545–574.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica 20*, 661–674.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society. Series B 70*(1), 53–71.

Ridout, M., J. Hinde, and C. G. B. Demétrio (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics 57*(1), 219–223.

Xiang, L., A. H. Lee, K. K. W. Yau, and G. J. McLachlan (2007). A score test for overdispersion in zero-inflated Poisson mixed regression model. *Statistics in Medicine 26*(7), 1608–1622.

Table 1. Simulation results for models $\mathcal{M}_1$: $\beta = (-.8, 1)$ and $\mathcal{M}_3$: $\beta = (-.8, 0)$

	percentage of immunes in the sample							
	0%		25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
Model $\mathcal{M}_1$								
100	-0.834	1.064	-0.773	1.114	-0.787	1.137	-0.750	0.917
	(0.258)	(0.301)	(0.583)	(0.412)	(0.825)	(0.603)	(0.921)	(0.858)
	[0.202]	[0.232]	[0.465]	[0.324]	[0.657]	[0.440]	[0.784]	[0.568]
		0.965*		0.109*		0.096*		0.121*
500	-0.807	1.012	-0.783	1.111	-0.788	1.129	-0.791	1.120
	(0.107)	(0.125)	(0.320)	(0.354)	(0.428)	(0.389)	(0.707)	(0.538)
	[0.085]	[0.099]	[0.264]	[0.227]	[0.352]	[0.270]	[0.603]	[0.407]
		1*		0.985*		0.85*		0.267*
1000	-0.801	1.004	-0.794	1.058	-0.798	1.060	-0.797	1.108
	(0.077)	(0.085)	(0.241)	(0.202)	(0.310)	(0.247)	(0.683)	(0.482)
	[0.062]	[0.068]	[0.201]	[0.147]	[0.253]	[0.178]	[0.569]	[0.354]
		1*		1*		1*		0.567*
1500	-0.805	1.003	-0.801	1.040	-0.799	1.040	-0.802	1.057
	(0.061)	(0.074)	(0.210)	(0.159)	(0.277)	(0.191)	(0.600)	(0.361)
	[0.048]	[0.059]	[0.176]	[0.119]	[0.228]	[0.141]	[0.493]	[0.276]
		1*		1*		1*		0.861*
Model $\mathcal{M}_3$								
100	-0.815	-0.001	-0.721	-0.007	-0.734	0.000	-0.746	-0.004
	(0.224)	(0.229)	(0.465)	(1.341)	(0.800)	(2.109)	(1.966)	(3.258)
	[0.177]	[0.179]	[0.377]	[0.762]	[0.636]	[1.111]	[1.516]	[1.715]
		0.052†		0.077†		0.069†		0.087†
500	-0.801	-0.001	-0.748	0.007	-0.750	0.001	-0.775	-0.006
	(0.097)	(0.099)	(0.280)	(0.415)	(0.520)	(0.469)	(1.209)	(0.711)
	[0.078]	[0.080]	[0.241]	[0.231]	[0.422]	[0.241]	[1.007]	[0.363]
		0.041†		0.058†		0.052†		0.057†
1000	-0.803	-0.001	-0.759	0.008	-0.763	0.005	-0.793	0.005
	(0.067)	(0.066)	(0.221)	(0.237)	(0.367)	(0.266)	(1.154)	(0.312)
	[0.053]	[0.053]	[0.182]	[0.137]	[0.299]	[0.140]	[0.911]	[0.175]
		0.042†		0.045†		0.037†		0.048†
1500	-0.801	0.000	-0.782	0.009	-0.784	0.003	-0.783	0.009
	(0.053)	(0.054)	(0.208)	(0.168)	(0.328)	(0.212)	(1.149)	(0.258)
	[0.042]	[0.043]	[0.178]	[0.099]	[0.267]	[0.102]	[0.901]	[0.144]
		0.051†		0.048†		0.027†		0.039†

Note: n: sample size. $(\cdot)$: root mean square error. $[\cdot]$: mean absolute error. *: empirical power (†: empirical size) of the Wald test at the level 5% for testing $H_0 : \beta_2 = 0$. For each percentage of immunes, the percentage of infected among the susceptibles is 30%. All results are based on 1500 replicates.

Table 2. Simulation results for models $\mathcal{M}_2$: $\beta = (1, .7)$ and $\mathcal{M}_4$: $\beta = (1, 0)$

	percentage of immunes in the sample							
	0%		25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
Model $\mathcal{M}_2$								
100	1.026	0.720	0.945	0.723	0.949	0.740	0.834	0.647
	(0.246)	(0.273)	(0.780)	(0.534)	(0.988)	(0.788)	(1.549)	(1.455)
	[0.196]	[0.215]	[0.655]	[0.376]	[0.829]	[0.555]	[1.326]	[0.933]
		0.746*		0.132*		0.118*		0.088*
500	1.003	0.712	1.098	0.717	1.112	0.721	0.840	0.652
	(0.107)	(0.115)	(0.651)	(0.247)	(0.672)	(0.279)	(0.969)	(0.534)
	[0.086]	[0.091]	[0.518]	[0.202]	[0.534]	[0.230]	[0.802]	[0.421]
		1*		0.503*		0.418*		0.168*
1000	1.003	0.707	1.078	0.711	1.096	0.719	0.842	0.657
	(0.071)	(0.082)	(0.590)	(0.215)	(0.571)	(0.224)	(0.796)	(0.439)
	[0.057]	[0.065]	[0.428]	[0.168]	[0.441]	[0.181]	[0.670]	[0.352]
		1*		0.779*		0.675*		0.205*
1500	1.001	0.701	1.035	0.705	1.069	0.709	0.887	0.655
	(0.064)	(0.065)	(0.450)	(0.163)	(0.466)	(0.177)	(0.604)	(0.312)
	[0.050]	[0.052]	[0.344]	[0.135]	[0.358]	[0.144]	[0.502]	[0.257]
		1*		0.986*		0.926*		0.300*
Model $\mathcal{M}_4$								
100	1.030	0.001	1.110	0.007	1.154	0.017	0.913	-0.003
	(0.233)	(0.234)	(0.852)	(0.969)	(1.211)	(1.347)	(1.775)	(1.640)
	[0.182]	[0.187]	[0.684]	[0.587]	[0.995]	[0.792]	[1.450]	[0.865]
		0.058†		0.072†		0.083†		0.066†
500	1.007	-0.005	1.105	0.020	1.123	0.054	0.915	-0.009
	(0.103)	(0.103)	(0.609)	(0.293)	(0.690)	(0.318)	(0.817)	(0.370)
	[0.081]	[0.082]	[0.492]	[0.180]	[0.562]	[0.208]	[0.614]	[0.215]
		0.046†		0.050†		0.063†		0.051†
1000	1.003	0.000	1.091	-0.003	1.101	0.033	0.934	-0.003
	(0.071)	(0.070)	(0.521)	(0.198)	(0.578)	(0.210)	(0.757)	(0.256)
	[0.057]	[0.055]	[0.437]	[0.125]	[0.455]	[0.135]	[0.600]	[0.142]
		0.051†		0.045†		0.042†		0.039†
1500	1.003	0.001	1.073	0.009	1.115	0.015	0.934	0.002
	(0.057)	(0.057)	(0.480)	(0.132)	(0.501)	(0.139)	(0.633)	(0.175)
	[0.046]	[0.046]	[0.392]	[0.087]	[0.400]	[0.104]	[0.521]	[0.109]
		0.042†		0.040†		0.046†		0.047†

Note: *: empirical power (†: empirical size) of the Wald test at the level 5% for testing $H_0 : \beta_2 = 0$. For each percentage of immunes, the percentage of infected among the susceptibles is 70%.

Table 3. "Naive" analysis of model $\mathcal{M}_1$: $\beta = (-.8, 1)$

	percentage of immunes in the sample					
	25%		50%		75%	
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$
100	-1.154	0.023	-1.632	0.017	-2.410	0.001
	(0.428)	(1.011)	(0.879)	(1.025)	(1.776)	(1.071)
	[0.365]	[0.977]	[0.833]	[0.983]	[1.610]	[1.003]
		0.049*		0.057*		0.052*
500	-1.128	0.087	-1.594	0.042	-2.305	0.002
	(0.344)	(0.915)	(0.803)	(0.963)	(1.513)	(1.010)
	[0.328]	[0.913]	[0.794]	[0.958]	[1.505]	[0.997]
		0.049*		0.051*		0.053*
1000	-1.131	0.059	-1.590	0.050	-2.297	0.033
	(0.338)	(0.941)	(0.795)	(0.952)	(1.501)	(0.970)
	[0.330]	[0.940]	[0.790]	[0.950]	[1.497]	[0.966]
		0.053*		0.051*		0.054*
1500	-1.127	0.050	-1.591	0.046	-2.302	0.039
	(0.332)	(0.953)	(0.794)	(0.955)	(1.504)	(0.962)
	[0.327]	[0.952]	[0.791]	[0.954]	[1.502]	[0.960]
		0.051*		0.050*		0.053*

Note: *: empirical power of the Wald test at the level 5% for testing $H_0 : \beta_2 = 0$. For each percentage of immunes, the percentage of infected among the susceptibles is 30%. In the "naive" analysis, every uninfected individual (*i.e.* $Y_i = 0$) is considered as susceptible.

Table 4. Dengue fever data analysis

	naive analysis		model (1)-(2)-(3)	
parameter	estimate	sd	estimate	sd
β_1	1.552	0.255	7.654	1.485
β_2	-0.055	0.007	-0.131	0.020
β_3	-0.813	0.207	-4.501	1.059
θ_1			0.497	0.159

Note: In the "naive" fitting of logit $\mathbb{P}(Y = 1|\text{age},\ \text{weight}) = \beta_1 + \beta_2\text{age} + \beta_3\text{weight}$, every uninfected individual is considered as susceptible. The final model (1)-(2)-(3) is given by logit $\mathbb{P}(Y = 1|\text{age},\ \text{weight}, S = 1) = \beta_1 + \beta_2\text{age} + \beta_3\text{weight}$ and logit $\mathbb{P}(S = 1) = \theta_1$.

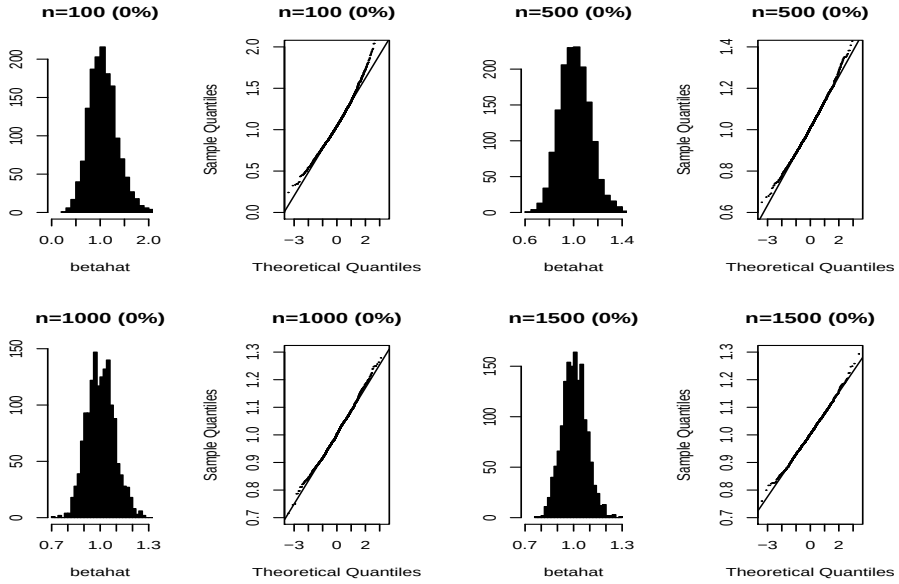

Figure 1. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with no immunes in the sample (the percentage of immunes is given in brackets). n is the sample size. All results are based on 1500 simulated datasets.

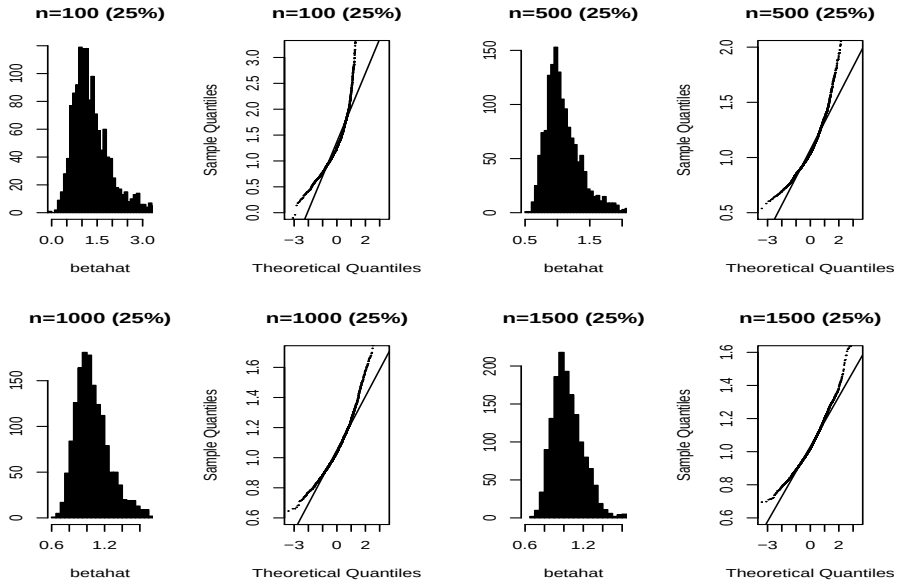

Figure 2. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with 25% of immunes.

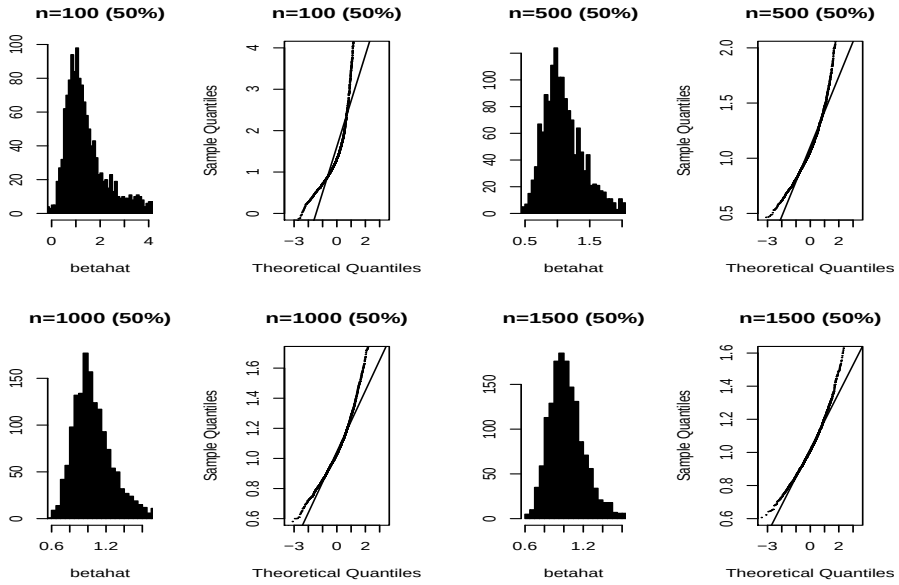

Figure 3. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with 50% of immunes.

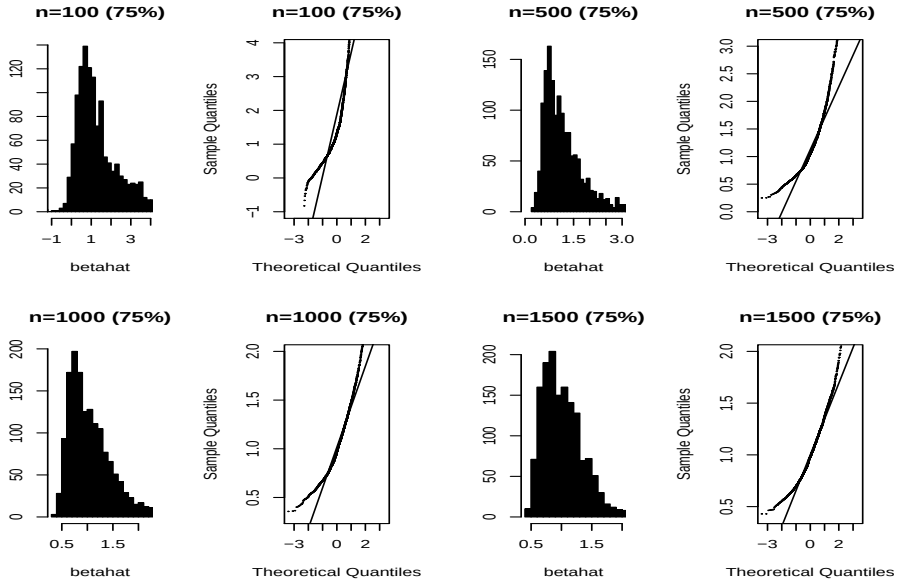

Figure 4. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with 75% of immunes.